Economics Letters 201 (2021) 109742

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Does monitoring deter future cheating? The case of external examiners in Italian schools☆

Marco Bertoni [a,b,*], Giorgio Brunello [a,b], Marco Alberto De Benedetto [c], Maria De Paola [c,b]

[a] University of Padova, Italy
[b] IZA, Germany
[c] University of Calabria, Italy

ARTICLE INFO

Article history:
Received 3 July 2020
Received in revised form 15 January 2021
Accepted 17 January 2021
Available online 22 February 2021

JEL classification:
H52
I2

Keywords:
Education
Testing
External monitoring
Short-run effects

ABSTRACT

We use the repeated random assignment of external examiners to schools in Italy to investigate whether the effect of external monitoring on test score manipulation persists over time. We find that this effect is still present in the tests taken one year after exposure to the examiners. In the second year after exposure, however, this effect disappears, suggesting that persistence is short lived.

© 2021 Elsevier B.V. All rights reserved.

0. Introduction

Appropriate incentives for law abidance require that individuals be monitored and sanctioned when they deviate from the prescribed behavior. Whether individuals who have been monitored (and eventually sanctioned) are more or less inclined to offend than other individuals is a central topic in the economics of law enforcement. Theoretically, agents can either reduce or increase their offending propensity, depending on whether they adjust upwards or downwards the probability of future punishment.

In this paper, we focus on the impact of past monitoring on current score manipulation in the standardized tests carried out by Italian schools and show that monitoring can affect subsequent compliance behavior even in a context where it is not accompanied by credible sanctioning. We investigate the effects of external monitoring in school *i* and years *t-2, t-1* and *t* on test scores in the same school in year *t*, using data on standardized math and literacy tests for the universe of 5th graders in Italian primary schools.

Our research design exploits the fact that, every year, external examiners in Italy are randomly allocated to groups of schools (called school institutes).[1] Examiners have the task of vigilating the entire test administration process, both by monitoring students taking the test and by supporting school staff in transcribing and transmitting the scores to the government agency in charge of test management.

We find evidence of short-term persistency: external monitoring in the school reduces the average percentage of correct answers and an index of cheating propensity both in the current and in the following year. While the effect of lagged monitoring on the percentage of correct answers is relatively small (−0.7 percent for math and −0.5 percent for literacy), the effect on cheating propensity is sizeable (−11.7 percent for math and −8.5 percent for literacy). After two years, however, the effects of having had an external examiner fade away completely.

1. Data

We consider the universe of 5th graders in Italian primary schools from 2014 to 2017. Our estimation sample consists of 22,984 observations in 6,790 school institutes. Every year the

☆ The authors are grateful to Erich Battistin, Michele De Nadai, Lorenzo Rocco and to the audience in Padova for comments and suggestions. Bertoni and Brunello acknowledge funding from a CARIPARO foundation "Starting Grant". The usual disclaimer applies.

* Correspondence to: Department of Economics and Management "Marco Fanno", University of Padova, Via del Santo 33, 35123 Padova, Italy.
E-mail addresses: marco.bertoni@unipd.it (M. Bertoni), giorgio.brunello@unipd.it (G. Brunello), marco_alberto.debenedetto@unical.it (M.A. De Benedetto), m.depaola@unical.it (M. De Paola).

[1] For the sake of brevity, we shall use the words "institutes" and "schools" interchangeably.

https://doi.org/10.1016/j.econlet.2021.109742
0165-1765/© 2021 Elsevier B.V. All rights reserved.

Italian agency in charge of running the tests (INVALSI) randomly selects a sample of institutes where a randomly selected subsample of one or two classes are subject to external monitoring. The sampling of institutes happens within region, and the probability of being sampled is proportional to the number of students enrolled. Samples are drawn independently every year.

In their analysis of monitoring in Italian schools, Angrist et al. (2017), show that the protocol for the randomization of external examiners is valid across institutes. They also show that the assignment of monitors to classes within institutes is suspect of deviations from randomness. Because of this, we use institutes as the unit of analysis, and define as treatment variables the presence of an examiner in the institute in year *t, t-1* and *t-2*.

For both math and literacy tests, we investigate the dynamic impact of examiners on the average percentage of correct answers (or score) given by each student. As discussed by Bertoni et al. (2013), manipulation is expected to reduce the variability of test results. Therefore, we also estimate the effects on the within-class standard deviation of the score, and on the class-level cheating propensity index computed by INVALSI.[2]

2. Empirical approach

We estimate the following empirical specification:

$$\begin{aligned} y_{irt} = {} & \alpha + \beta_1 Monitored_{irt} + \beta_2 Monitored_{irt-1} + \beta_3 Monitored_{irt-2} \\ & + \delta_{1rt} Size_{irt} + \delta_{2rt} Size_{irt-1} + \delta_{3rt} Size_{irt-2} + \mu_{rt} \\ & + \gamma_1 X_{irt} + \gamma_2 W_{irt-1} + \gamma_3 Z_{irt-2} + \varepsilon_{irt} \end{aligned} \quad (1)$$

where the indices *i, r* and *t* are for school, region and year; *y* is the outcome variable — measured in year *t*; $Monitored_{irt}$, $Monitored_{irt-1}$ and $Monitored_{irt-2}$ are binary variables equal to 1 if external examiners proctored the test in school *i* in years *t*, *t-1* and *t-2*, and to 0 otherwise.

If the current assignment of an external monitor reduces score manipulation (or cheating), coefficient β_1 should be negative for all our outcomes except the standard deviation of test scores (for which it should be positive). On the other hand, if the assignment of an external monitor in year *t-1* or *t-2* has no persistent effect on current outcomes, coefficients β_2 and/or β_3 should be equal to zero.

We take into account the INVALSI randomization protocol by including in the specification both region-by-year dummies (μ_{rt}) and their interactions with school size in year *t, t-1* and *t-2* ($\delta_{1rt} Size_{irt}$, $\delta_{2rt} Size_{irt-1}$ and $\delta_{3rt} Size_{irt-2}$).

In addition, X_{irt} is a vector of control variables which includes the share of male and immigrant students; the share of mothers and fathers with an elementary, middle, high-school diploma and a degree; the share of students who attended pre-primary schools; the share of students following a full-day schedule and the share of irregular students. We further include in vector X_{irt} the share of missing values for each of the covariates described above. The vectors W_{irt-1} and Z_{irt-2} contain the same variables included in the vector X_{irt}, but measured in year *t-1* and *t-2*, respectively. Finally, ε_{irt} is an error term that we allow to be clustered by school. As we are considering several treatment timings and outcomes, we also compute stepdown p-values correcting for the effect of multiple hypothesis testing as proposed by Romano and Wolf (2005), using 250 bootstrap replications.

[2] This index (similar to the one used by Angrist et al., 2017) is estimated as the degree of belongingness to an "outlier cluster" determined by the application of a fuzzy clustering algorithm to the following features of the class-level score distribution: the mean score, the standard deviation of the mean score, the non-response rate, the Herfindahl index of answers' homogeneity. For a detailed description see Quintano et al. (2009).

3. Results

Our results for math and literacy are reported in Table 1. Consistent with the previous literature, we find that the percentage of correct answers in schools where an external examiner was present at the test taken in year *t* is 4.1 percent lower for literacy and 5.4 percent lower for math than in schools that did not have an external examiner — see columns (1) and (4).

If the presence of an external examiner had only a temporary effect of average school test scores, having had an examiner in year *t-1* or *t-2* should have no effect on test scores in year *t*. Yet we find that schools which had an external examiner during the test taken at *t-1* experience a statistically significant reduction in the percent of correct answers in the test taken in year *t*, ranging from 0.5 percent for literacy to 0.7 percent for math. This effect is roughly 1/7 of the current examiner's effect. Persistency, however, is short-lived: the binary treatment in year *t-2* does not produce a statistically significant effect on test scores at time *t*, implying that over time schools revert to their original behavior.

In addition, as shown in columns (2) and (5), there is a significant and positive effect of monitoring in year *t* on the standard deviation of scores in year *t*. However, the impact of having had an external examiner in the school in year *t-1* is only significant for the math test, and the effect of monitoring in year *t-2* is always very close to zero.

In columns (3) and (6) we turn our attention to the INVALSI cheating index. We find that schools being monitored in year *t* show a large reduction in the cheating index — ranging between 43 and 50 percent. Having been monitored in year *t-1* also reduces the index by 8.4 percent for literacy and by 11.7 percent for math. No effect is found instead for monitoring in year *t-2*.[3]

We also investigate whether the impact of external monitoring varies across macro areas according to their level of social capital, and show that short-lived persistency is present in local areas with high social capital – in the Northern and Central regions of Italy – and absent in the South, where social capital is much lower (see Tables A2 and A3 in the Appendix). As long as areas of Italy endowed with high social capital also share a self-image prescribing not to cheat, these results suggest that monitors may make this identity more salient, promoting conformism even in the year after the examiners visited the school.[4]

4. Conclusions

A well-known problem with testing is score manipulation, which happens both in low and high stakes tests, and undermines both the reliability of results and the possibility of using them to compare schools and countries and support accountability policies. The existing empirical evidence shows that external monitoring is effective in reducing manipulation. The relevant literature, however, has focused exclusively on the immediate incapacitation effects of monitoring, with the implicit assumption that these effects vanish once external invigilators leave the school.

In this paper, we have questioned this assumption by investigating whether the presence of external examiners can also impact future test scores. Using the repeated random assignment of external examiners to Italian primary schools, we have found that external monitoring reduces average test scores and cheating not only currently but also in the year after its implementation. After two years, however, the effect vanishes.

[3] Since controls in X_{irt}, W_{irt-1} and Z_{irt-2} might be highly correlated, we present regression results without these controls in Table A1 in Appendix. Findings are similar to those reported in Table 1.

[4] See the discussion paper version of this paper – Bertoni et al. (2019) – for further details.

Table 1
The effects of external monitoring on test scores. Math and literacy — 5th graders.

Outcome variable	(1) Mean – math	(2) Standard deviation – math	(3) Cheating index – math	(4) Mean – literacy	(5) Standard deviation – literacy	(6) Cheating index – literacy
Monitored in year *t*	−3.372***	0.626***	−0.020***	−2.690***	0.718***	−0.020***
	(0.195)	(0.054)	(0.001)	(0.155)	(0.054)	(0.001)
	[<0.01]	[<0.01]	[<0.01]	[<0.01]	[<0.01]	[<0.01]
Monitored in year *t-1*	−0.440*	0.130*	−0.005***	−0.312*	0.024	−0.003*
	(0.204)	(0.059)	(0.002)	(0.163)	(0.057)	(0.001)
	[0.08]	[0.08]	[<0.01]	[0.09]	[0.67]	[0.06]
Monitored in year *t-2*	0.172	−0.001	0.001	0.135	0.009	0.001
	(0.196)	(0.059)	(0.002)	(0.158)	(0.058)	(0.001)
	[0.82]	[0.99]	[0.83]	[0.82]	[0.97]	[0.90]
Observations	22,984	22,984	22,984	22,984	22,984	22,984
Mean for control group at *t*	62.11	15.17	0.046	64.27	15.07	0.040
Mean for control group at *t-1*	61.85	15.22	0.045	64.07	15.13	0.038
Mean for control group at *t-2*	61.69	15.25	0.045	63.99	15.14	0.038
% change for monitored at *t*	−0.054	0.041	−0.426	−0.041	0.047	−0.499
% change for monitored at *t-1*	−0.007	0.008	−0.117	−0.004	0.001	−0.084
% change for monitored at *t-2*	0.002	−0.001	0.028	0.002	0.001	0.019

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrollment) and the other controls in vectors X, W and Z. Standard errors clustered by school in parentheses. Romano–Wolf stepdown p-values correcting for the effect of multiple hypothesis testing (250 replications) in brackets. All the 18 coefficients reported in the Table are included in the multiple test. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect by the mean outcome for the control group. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence according to the Romano–Wolf p-values.

Our findings have implications for the design of policies using external monitors to deter cheating in school tests. In the areas where the effects of external monitoring persist into the next year (the North and Centre of Italy in our study), the frequency of interventions could be reduced (for instance every two years) freeing up scarce resources to intensify yearly monitoring in the areas where social capital is low (the South of Italy). By so doing, the reduction in overall manipulation would be higher.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2021.109742.

References

Angrist, J.D., Battistin, E., Vuri, D., 2017. In a small moment: Class size and moral hazard in the Italian Mezzogiorno. Am. Econ. J.: Appl. Econ. 9 (4), 216–249.
Bertoni, M., Brunello, G., De Benedetto, M.A., De Paola, M., 2019. External monitors and score manipulation in Italian schools: symptomatic treatment or cure? IZA Discussion Paper 12591.
Bertoni, M., Brunello, G., Rocco, L., 2013. When the cat is near, the mice won't play: The effect of external examiners in Italian schools. J. Publ. Econ. 104, 65–77.
Quintano, C., Castellano, R., Longobardi, S., 2009. A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of outliers on assessment test scores. Stat. Appl. 7, 149–171.
Romano, J.P., Wolf, M., 2005. Exact and approximate stepdown methods for multiple hypothesis testing. J. Amer. Statist. Assoc. 100 (469), 94–108.